Assignment #5

Jingye Wang

October 5, 2016

```
setwd("~/Dropbox/WUSTL third/Multilevel Modeling for Quantitative
Research/assignment/5")
.libPaths("/Library/Frameworks/R.framework/Versions/3.3/Resources/library")
```

Ch14.5

Multilevel logistic regression with non-nested groupings:the folder speed.dating contains data from an experiment on a few hundred students that randomly assigned each participant to 10 short dates with participants of the opposite sex (Fisman et al., 2006). For each date, each person recorded several subjective numerical ratings of the other person (attractiveness, compatibility, and some other characteristics) and also wrote down whether he or she would like to meet the other person again.

Part A

Fit a classical logistic regression predicting Pr(yij = 1) given person i's 6 ratings of person j. Discuss the importance of attractiveness, compatibility, and so forth in this predictive model.

import data

```
date <- read.csv('Speed Dating Data.csv', header = T)
date<-date[,c("dec","iid","pid","attr","sinc", "intel","fun", "amb","shar")]
```

fit the model

```
model_1 <- glm(dec~ attr + sinc + intel + fun + amb + shar, data=date,
family="binomial")
exp(cbind(OR=coef(model_1), confint(model_1)))
```

```
## Waiting for profiling to be done...
```

```
##                      OR       2.5 %      97.5 %
## (Intercept) 0.004994684 0.003446987 0.007184786
## attr        1.729884111 1.658715690 1.805371080
## sinc        0.894347666 0.851455930 0.939288468
## intel       1.028358556 0.968958145 1.091354639
## fun         1.303704269 1.244638214 1.366036261
## amb         0.845952967 0.807445344 0.886023718
## shar        1.307803491 1.261513646 1.356236187
```

```
summary(model_1)
```

```
##
## Call:
## glm(formula = dec ~ attr + sinc + intel + fun + amb + shar, family = "binomial",
##     data = date)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5147  -0.8377  -0.3071   0.8583   3.3832
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.29938    0.18735 -28.287  < 2e-16 ***
## attr         0.54805    0.02161  25.361  < 2e-16 ***
## sinc        -0.11166    0.02504  -4.459 8.23e-06 ***
## intel        0.02796    0.03034   0.922    0.357
## fun          0.26521    0.02374  11.172  < 2e-16 ***
## amb         -0.16729    0.02369  -7.062 1.64e-12 ***
## shar         0.26835    0.01847  14.531  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9626.0  on 7039  degrees of freedom
## Residual deviance: 7160.4  on 7033  degrees of freedom
##   (1338 observations deleted due to missingness)
## AIC: 7174.4
##
## Number of Fisher Scoring iterations: 5
```

Discussion: Based on this model, we can find that other five variables are statistically significant indicators of the decision making except for the intelligence. According to the odds ratio of each variable, attractive is the most positive factor of the outcome. Fun and shared interests are second and third positive factors, while ambitious and sincere are negtive factors. Additionally, intelligence is not important for people's dertermine of whether to date again.

Part B

Expand this model to allow varying intercepts for the persons making the evaluation; that is, some people are more likely than others to want to meet someone again. Discuss the fitted model.

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
group_model_1_1 <- glmer(dec~ attr + sinc + intel + fun + amb + shar + (1 | iid), data=d
ate, family="binomial")
summary(group_model_1_1)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: dec ~ attr + sinc + intel + fun + amb + shar + (1 | iid)
##    Data: date
##
##      AIC      BIC   logLik deviance df.resid
##   5712.5   5767.4  -2848.3   5696.5     7032
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -26.2883  -0.3523  -0.0541   0.3449  13.7737
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  iid    (Intercept) 5.419    2.328
## Number of obs: 7040, groups:  iid, 537
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -12.77069    0.46988 -27.179  < 2e-16 ***
## attr          0.90768    0.03533  25.690  < 2e-16 ***
## sinc          0.05587    0.03732   1.497   0.1344
## intel         0.17449    0.04390   3.974 7.06e-05 ***
## fun           0.45052    0.03404  13.234  < 2e-16 ***
## amb          -0.05924    0.03434  -1.725   0.0845 .
## shar          0.40822    0.02889  14.132  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) attr   sinc   intel  fun    amb
## attr  -0.538
## sinc  -0.294  0.033
## intel -0.316  0.048 -0.383
## fun   -0.241 -0.035 -0.096 -0.047
## amb   -0.184 -0.036  0.040 -0.306 -0.151
## shar  -0.217  0.081 -0.026 -0.024 -0.188 -0.106
## convergence code: 0
## Model failed to converge with max|grad| = 0.0108786 (tol = 0.001, component 1)
```

Discussion: Based on this model, we can find that four variables are statistically significant indicators of the decision making. Intelligence and ambitious are not statistically significant, while ambitious is a negtive factor. Additionally, based on the random effects, people’s dertermine of whether to date again vary from different people.

Part C

Expand further to allow varying intercepts for the persons being rated. Discuss the fitted model.

```
group_model_1_2 <- glmer(dec~ attr + sinc + intel + fun + amb + shar + (1 | iid) + (1 | pi
, data=date, family="binomial")
summary(group_model_1_2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: dec ~ attr + sinc + intel + fun + amb + shar + (1 | iid) + (1 |
##     pid)
##    Data: date
##
##      AIC      BIC   logLik deviance df.resid
##   5695.8   5757.6  -2838.9   5677.8     7022
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -24.1258  -0.3375  -0.0500   0.3337  13.5052
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  pid    (Intercept) 0.2087   0.4569
##  iid    (Intercept) 5.6011   2.3667
## Number of obs: 7031, groups:  pid, 551; iid, 537
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -12.83839    0.48505 -26.468  < 2e-16 ***
## attr          0.90142    0.03701  24.354  < 2e-16 ***
## sinc          0.06889    0.03887   1.772 0.076376 .
## intel         0.16435    0.04563   3.602 0.000316 ***
## fun           0.46410    0.03578  12.970  < 2e-16 ***
## amb          -0.06650    0.03567  -1.864 0.062297 .
## shar          0.41520    0.02986  13.906  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) attr   sinc   intel  fun    amb
## attr  -0.523
## sinc  -0.296  0.024
## intel -0.309  0.036 -0.382
## fun   -0.246 -0.042 -0.086 -0.059
## amb   -0.179 -0.034  0.029 -0.300 -0.152
## shar  -0.219  0.070 -0.017 -0.022 -0.176 -0.114
## convergence code: 0
## Model failed to converge with max|grad| = 0.00963154 (tol = 0.001, component 1)
```

Discussion: Based on this model, we can find that four variables are statistically significant indicators of the decision making. Intelligence and ambitious are not statistically significant, while ambitious is a negtive factor. Additionally, based on the random effects, people's dertermine of whether to date again vary from different people and people being rated. Additionally, the group level differences between people are larger than the differences between people being rated.

Ch14.6

Varying-intercept, varying-slope logistic regression: continuing with the speed dating example from the previous exercise, you will now fit some models that allow the coefficients for attractiveness, compatibility, and the other attributes to vary by person.

Part A

Fit a no-pooling model: for each person i, fit a logistic regression to the data yij for the 10 persons j whom he or she rated, using as predictors the 6 ratings rij1 , . . . , rij6 . (Hint: with 10 data points and 6 predictors, this model is difficult to fit. You will need to simplify it in some way to get reasonable fits.)

```
date$ave_score <- apply(date[, -c(1:3)], 1, function(x) mean(x, na.rm = T))
model_nopool <- lmList(dec ~ ave_score|iid, data = date, family = 'binomial')
```

Part B

Fit a multilevel model, allowing the intercept and the coefficients for the 6 ratings to vary by the rater i.

```
model_2 <- glmer(dec~ ave_score + (1 + ave_score | iid), data=date, family="binomial"(li
nk = "logit"))
summary(model_2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: dec ~ ave_score + (1 + ave_score | iid)
##    Data: date
## 
##      AIC      BIC   logLik deviance df.resid 
##   7420.2   7455.2  -3705.1   7410.2     8181 
## 
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -6.7075 -0.4188 -0.0834  0.4123  6.2202 
## 
## Random effects:
##  Groups Name        Variance Std.Dev. Corr 
##  iid    (Intercept) 31.8718  5.6455        
##         ave_score    0.6088  0.7803   -0.92
## Number of obs: 8186, groups:  iid, 551
## 
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -14.68464    0.52769  -27.83   <2e-16 ***
## ave_score     2.12460    0.07589   28.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Correlation of Fixed Effects:
##           (Intr)
## ave_score -0.979
```

Discussion: Based on this model, we can find that the std.dev of intercept (5.6455) are much larger than the std.dev of ave_score (0.7803), which means that intercept is more suitble to build a multi-level model than ave_score. Additionally, both of them are statistically significant indicators for the decision making.

Part C

Compare the inferences from the multilevel model in (b) to the no-pooling model in (a) and the complete-pooling model from part (a) of the previous exercise.

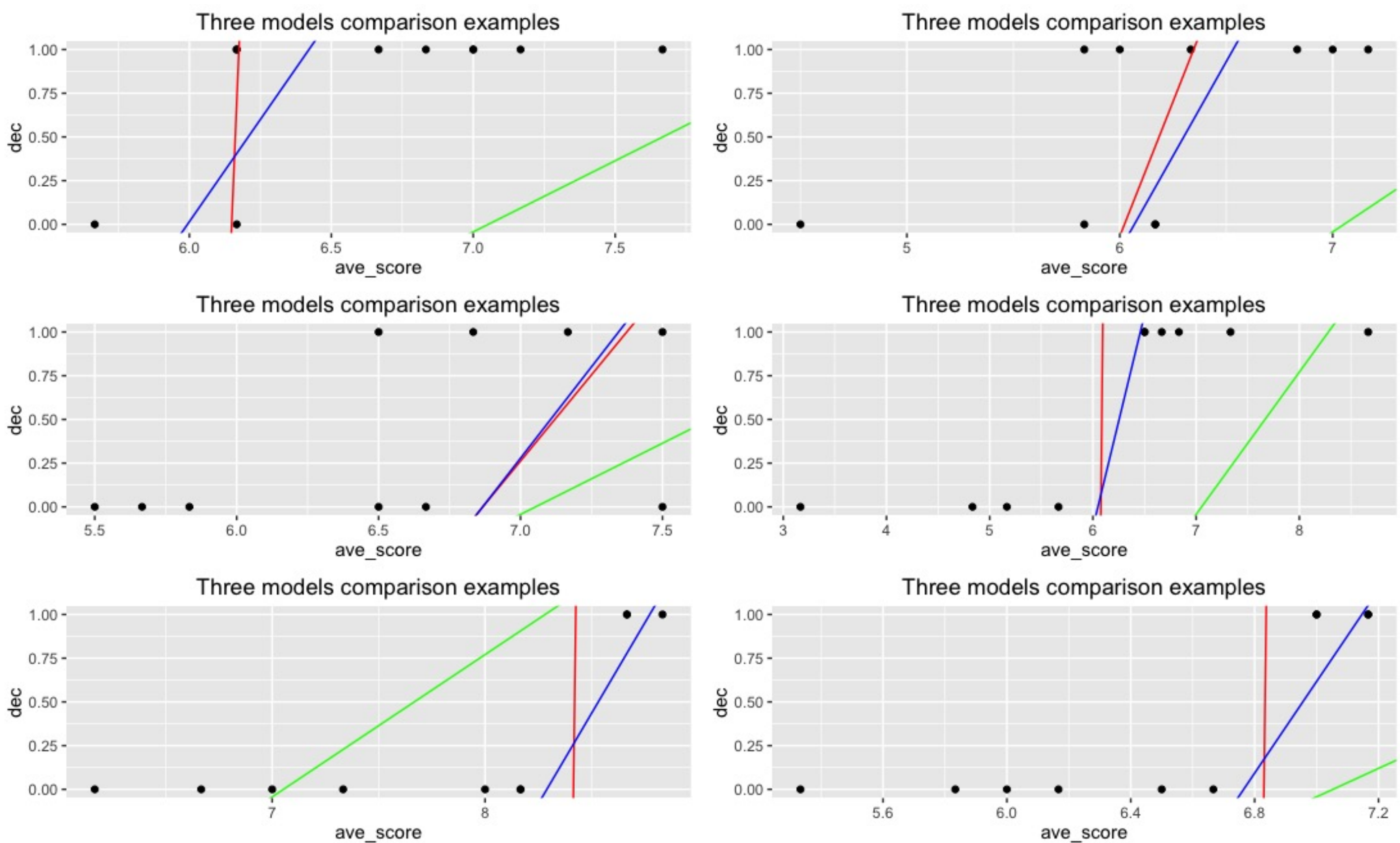
fit the complete-pooling model

```
model_3 <- glm(dec ~ ave_score, data = date, family = 'binomial')
summary(model_3)
```

```
##
## Call:
## glm(formula = dec ~ ave_score, family = "binomial", data = date)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2295  -0.9408  -0.5047   1.0273   2.6328
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.73979    0.15783  -36.37   <2e-16 ***
## ave_score    0.81381    0.02296   35.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11182.7  on 8185  degrees of freedom
## Residual deviance:  9426.5  on 8184  degrees of freedom
##   (192 observations deleted due to missingness)
## AIC: 9430.5
##
## Number of Fisher Scoring iterations: 4
```

```
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
  library(grid)

  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)

  numPlots = length(plots)

  # If layout is NULL, then use 'cols' to determine layout
  if (is.null(layout)) {
    # Make the panel
    # ncol: Number of columns of plots
    # nrow: Number of rows needed, calculated from # of cols
    layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                     ncol = cols, nrow = ceiling(numPlots/cols))
  }

  if (numPlots==1) {
    print(plots[[1]])

  } else {
    # Set up the page
    grid.newpage()
    pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

    # Make each plot, in the correct location
    for (i in 1:numPlots) {
      # Get the i,j matrix positions of the regions that contain this subplot
      matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

      print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                      layout.pos.col = matchidx$col))
    }
  }
}
```

```
library(ggplot2)
coef_df_1 <- data.frame(coef(model_nopool))
colnames(coef_df_1) <- c('Inter', 'Slope')
coef_df_2 <- data.frame(coef(model_2)$iid)
colnames(coef_df_2) <- c('Inter', 'Slope')
coef_df_3 <- data.frame(t(coef(model_3)))
colnames(coef_df_3) <- c('Inter', 'Slope')
plot_list = list()
for (i in 1:10){
p <- ggplot(data = date[date$iid==i, ], aes(ave_score, dec)) + geom_point() + geom_ablin
e(slope = coef_df_1[i, 2], intercept = coef_df_1[i, 1], colour = 'red') + geom_abline(sl
ope = coef_df_2[i, 2], intercept = coef_df_2[i, 1], colour = 'blue') + geom_abline(slope
 = coef_df_3[1, 2], intercept = coef_df_3[1, 1], colour = 'green') + ggtitle('Three mode
ls comparison examples')
plot_list[[i]] = p
}
multiplot(plot_list[[1]],plot_list[[2]],plot_list[[4]],plot_list[[5]],plot_list[[6]],plo
t_list[[7]],cols=2)
```



Discussion: Based on examples of these three models comparison, the no-pooling lines and the multi-level lines are relatively close with each other. However, the complete-pooling model are quite different from other two models.

Ch15.1

Multilevel ordered logit: using the National Election Study data from the year 2000 (data available in the folder nes), set up an ordered logistic regression predicting the response to the question on vote intention (0 = Gore, 1 = no opinion or other, 2 = Bush), given the predictors shown in Figure 5.4 on page 84, and with varying intercepts for states. (You will fit the model using Bugs in Exercise 17.10.)

- $y = \begin{cases} 0 \ if \ Z_i < 0 \\ 0 \ if \ Z_i \epsilon \ (0, c) \\ 0 \ if \ Z_i > c \end{cases}$
- $Zi = \alpha_{j[i]} + \beta x_i + \varepsilon_i$
- $j \ \epsilon \ \{State1, State2, \dots\}$

Ch15.2

Using the same data as the previous exercise

Part A

Formulate a model to predict party identification (which is on a five-point scale) using ideology and demographics with a multilevel ordered categorical model allowing both the intercept and the coefficient on ideology to vary over state.

- $y = \begin{cases} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{cases}$
- $Z_i = \alpha_{1j[i]} + \alpha_{2j[i]} + \cdots + \beta_{1j[i]} x_1 + \beta_{2j[i]} x_2 + \cdots + \varepsilon_i$

Part B

Fit the model using lmer() and discuss your results.

```
library(foreign)
nes_data <- read.dta('nes5200_processed_voters_realideo.dta')
nes_data$party_num <- as.numeric(gsub('([0-9]).*', '\\1', nes_data$partyid7))
nes_2000 <- subset(nes_data, year == 2000)
myvars1<-c("party_num","ideo7","state", "age")
nes_2000<-na.omit(nes_2000[myvars1])
nes_2000$state <-as.factor(nes_2000$state)

model_4 <- lmer(party_num ~ ideo7 + age + (1 + ideo7 + age | state), data = nes_2000)
summary(model_4)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: party_num ~ ideo7 + age + (1 + ideo7 + age | state)
##    Data: nes_2000
##
## REML criterion at convergence: 2764.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4206 -0.6944  0.0103  0.7599  2.3100
##
## Random effects:
##  Groups   Name                                 Variance  Std.Dev.  Corr
##  state    (Intercept)                          6.385e-09 0.0000799
##           ideo72. liberal                      1.380e+00 1.1749442  0.14
##           ideo73. slightly liberal             1.339e+00 1.1572163  0.13
##           ideo74. moderate, middle of the road 5.468e-01 0.7394374 -0.01
##           ideo75. slightly conservative        2.173e+00 1.4740709  0.16
##           ideo76. conservative                 1.024e+00 1.0120123 -0.07
##           ideo77. extremely conservative       3.349e+00 1.8299052  0.17
##           age                                  2.844e-04 0.0168645 -0.14
##  Residual                                      3.209e+00 1.7913098
##
##
##
##   0.32
##   0.66  0.65
##   0.34  0.34  0.51
##   0.20  0.45  0.61  0.65
##   0.25  0.26  0.33  0.37  0.41
##  -0.53 -0.68 -0.82 -0.87 -0.74 -0.54
##
## Number of obs: 667, groups:  state, 46
##
## Fixed effects:
##                                       Estimate Std. Error t value
## (Intercept)                           1.790894   0.709304   2.525
## ideo72. liberal                       0.507917   0.738952   0.687
## ideo73. slightly liberal              1.160999   0.745073   1.558
## ideo74. moderate, middle of the road  1.960481   0.697466   2.811
## ideo75. slightly conservative         2.974448   0.745660   3.989
## ideo76. conservative                  2.985670   0.709768   4.207
## ideo77. extremely conservative        4.163285   0.876002   4.753
## age                                  -0.003313   0.005625  -0.589
##
## Correlation of Fixed Effects:
##             (Intr) id72.l i73.sl immotr i75.sc id76.c i77.ec
## ideo72.lbrl -0.869
## id73.slghtl -0.853  0.845
## id74.m,motr -0.916  0.913  0.903
## id75.slghtc -0.840  0.848  0.839  0.896
## id76.cnsrvt -0.896  0.876  0.885  0.941  0.906
## id77.extrmc -0.719  0.721  0.717  0.757  0.730  0.762
## age         -0.275 -0.059 -0.101 -0.071 -0.184 -0.096 -0.117
```

```
## convergence code: 1
## unable to evaluate scaled gradient
## Model failed to converge: degenerate  Hessian with 3 negative eigenvalues
## maxfun < 10 * length(par)^2 is not recommended.
```

Discussion: Based on this model, compare to the std.dev of residual (1.79483), age (0.01769) and intercept (0.06990) are not suitble for a multi-level model. For fixed effects, ideo7 is a positive predictor, while age is a negative predictor.